# Submission
# 2016 National Research Infrastructure Roadmap
# Capability Issues Paper

| Name | Nick Thieberger |
|------|-----------------|
| Title/role | Director |
| Organisation | Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) |

**General Questions**

**Question 5: Should research workforce skills be considered a research infrastructure issue?**

Most research infrastructure will not work without trained staff. A major problem with short-term funding (e.g. on an annual basis) is that experienced staff are not retained. Funds should be assured over longer periods to help build confidence in the infrastructure developed and also build a skilled researcher base.

**Question 6: How can national research infrastructure assist in training and skills development?**

In our experience, building a major repository has entailed running training in data management, as well as in discipline-specific methods for creating well-formed research data that can be archived. The two are integral to each other, a repository does not want to be renaming and reformatting files for each deposit if it can avoid it, and the researcher wants to be able to provide a properly described collection that reflects their own knowledge of the content. They can then be confident that they will not be dealing with requests to access their material in a continual adhoc way, but that the repository will allow them to act in an ethical way by making their material available with clear access conditions and rights statements.

**Question 7: What responsibility should research institutions have in supporting the development of infrastructure ready researchers and technical specialists?**

It is central to good research practice that all researchers understand the nature of data management appropriate for their discipline. This understanding can be reached via discipline-specific training or is able to be dealt with in more generic data-management courses provided by their institution. We have to keep in mind in HASS that our graduates will be running all kinds of public entities, perhaps in the public service or in private industry, and they all need to have far better understanding of the role of technology than they are currently being provided in normal research training efforts.

**Question 8: What principles should be applied for access to national research infrastructure, and are there situations when these should not apply?**

We will take national research infrastructure to include the National Library of Australia and similar institutions (NAA, NFSA) (as recognised at 8.1.3. of the 2016 Roadmap). These need to be funded at a higher level as they are the research base for the majority of research carried out in Australia, especially that conducted by researchers outside of universities.

Other national research infrastructure, like ANDS and storage capacity (previously known as RDS) needs to be flexible enough to understand the range of disciplinary requirements, and not just provide petabyte solutions tailored to astrophysics. ANDS has been responsive to a diverse community of users, it would seem appropriate for it to actually become a data service in future.

**Question 10: What financing models should the Government consider to support investment in national research infrastructure?**

The government should be able to fund research in Australia to a much higher level than it currently does and it should regard this funding as an investment both in the research material that is produced, but also in the skills developed in Australian graduates who will go on to become managers of public and private enterprise. The lack of understanding of basic data management practices has cost enormous amounts in wasted investment when IT companies are able to walk away from a failed project because there was no proper expertise inhouse to evaluate and manage the project. This expertise can be developed during training in Digital Humanities methods, for example, to a greater extent than it is at present.

It is vitally important that research data curation and storage should be provided in the longterm for all research data outputs. PARADISEC has established a system for curation of records in the many small languages of the world. This is a project that has so far lasted 13 years and keeps growing in relevance over time.  We have received the European 'Data Seal of Approval' for our collection( http://www.datasealofapproval.org/en/community/) , and we are members of international consortia such as  DELAMAN  (Digital Endangered Languages and Music Archives Network, _____ The metadata catalogue is compliant with OLAC (Open Languages Archives Community, http://www.language-archives.org/).

For data considered of state/national importance there needs to be a long-term funding strategy so that guarantees of data migration and data preservation can be made. We should look to European models of data archives, where guarantees of 50 years are given. The move to a so-called 'user-pays' model for data storage is shortsighted. In the past we have seen projects that have had large investments by the tax-payer, both through universities and through government departments (such as departments of education) commissioning research. The research materials may be of high value, but those materials are now lost. Data storage needs to be provided for free (perhaps on the same 'merit-allocation' basis as it has been offered in the past) in order to safeguard the investment that was made in the original research.

## Data transformation into research data

While there is an understandable emphasis on massive data sets in the Roadmap draft document (given that the overwhelming majority of its focus is on STEM), it should be understood that most research (in terms of numbers of people doing and receiving the research) being carried out in Australia produces very small, often highly crafted data sets. In HASS, an individual or small team will work on a particular set of material, perhaps a historical document or an author's life work, or a set of recordings made in an indigenous language. These are not massive datasets, but are finely curated and enriched datasets. This work needs to be captured, with the added value (e.g. annotations) that has to be kept together with the primary data. Training needs to be provided so that, for example, a researcher taking images of manuscripts for their own research would think of their collection of digital images as a resource that could be re-used by others and so will conform to appropriate standards that they have been trained in, for data formats and naming conventions of the files. They will have a suitable repository available into which to deposit these files, and the metadata will be picked up by ANDS so that others can discover the existence of collection.

## Fundraising for research data futures

How the government raises funds is a political issue, but since you ask, it would seem that leadership by the government to publicise its recognition that research strengthens Australia's future would be part of justifying to the electorate the increase in funding. Governments choose where to invest funds, and, while obsolete defence equipment seems to be an easy choice that costs huge amounts, a fraction of that funding could be used to support the NLA, NFSA and NAA who house Australia's cultural wealth, to fund university research and to provide data storage and to recognise the asset that this material represents.

**Question 11: When should capabilities be expected to address standard and accreditation requirements?**

Always

**Question 12: Are there international or global models that represent best practice for national research infrastructure that could be considered?**

[No answer provided]

**Question 13: In considering whole of life investment including decommissioning or defunding for national research infrastructure are there examples domestic or international that should be examined?**

[No answer provided]

**Question 14: Are there alternative financing options, including international models that the Government could consider to support investment in national research infrastructure?**

There are a number of European models for funding HASS projects, CLARIN, Dariah are recent examples, but the German experience of privatising state enterprises (like Volkswagen) and putting the revenue into a Foundation that funds ongoing research programs independent of government is a good model.

## Understanding Cultures and Communities

**Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?**

There are many more examples of HASS datasets that should be recognised in this document. While STEM is divided into a number of what you call 'capabilities' all HASS disciplines are listed in just one. There is a great deal of HASS activity and it is a great disservice (not to mention a great disrespect) to the researchers involved in that work that you do not consider them in this draft. For example: AustLit; ADB; AusStage; Founders and Survivors; Bright Sparcs; South Seas; the Living Archive of Aboriginal Languages; PARADISEC;

It is also strange that university research is not listed in the present draft as it is centrally important to this Capability. While galleries, libraries and museums (GLAM) need more resources, most research that produces data that needs curation is produced by university-based researchers. There is thus a need for national coordination of the curation of primary research data from both the GLAM and university sectors.

This group should also have considered the report by Turner and Brass (2014:1) who say "The evidence presented in this report demonstrates that Australia has a strong and resilient HASS sector that makes a major contribution to the national higher education system, to the national research and innovation system, and to preparing our citizens for participation in the workforce. The vast majority of tertiary enrolments are in HASS programmes."

Turner, G., and Brass, K. (2014) Mapping the Humanities, Arts and Social Sciences in Australia. Australian Academy of the Humanities, Canberra.

**Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?**

[No answer provided]

**Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area?**

Yes, quite a lot. In addition to those laid out earlier, it would be worth looking at the 2012 Roadmap where there is a much a better consideration of this capability than that produced by the present working group. Some extracts are re-presented here:

> The unmet demand for infrastructure investment, the growing emphasis on collaborative multidisciplinary research and the continued investment being made in the United States and Europe in HASS-orientated research infrastructure means that the sector now requires substantial investment to bring supporting infrastructure to a standard which will enable a critical degree of multidisciplinary integration and enable continued significant international contributions.

> While the conceptualisation and scope of the 2008 Capability remains relevant to present initiatives, new technological possibilities now allow us to better define immediate needs and longer-term directions.  Access to diverse sources of data in an integrated and cost-effective manner is a key priority.

> The proposed national eResearch facility would provide a distributed national online environment and the tools needed for interacting and collaborating, and for generating, discovering, accessing, working with and publishing data, regardless of physical location or format.  Data in this sector exists in a plethora of formats, many of which are currently very difficult to align for the purpose of meaningful analysis.

> Much data of interest to researchers engaged with understanding cultures and communities remains in individual repositories in analogue form and in some cases this may necessitate transfer to appropriate digital formats.  A one-size-fits-all approach cannot deal adequately with this level of complexity.  While we can learn from the experience of existing Capabilities, with some elements adapted for our use, addressing the research needs of this sector will require purpose-designed infrastructure.

> European Commission, Directorate-General for Research, Synergies Between FP7 and Structural Funds for Research Infrastructures, Report, 29 September

2010 http://ec.europa.eu/research/infrastructures/pdf/synergies-fp-sfmappingesfriprojects.pdf

Inspiring Excellence Research Infrastructures and the Europe 2020 Strategy, by Carlo Rizzuto, Chair, ESFRI (European Strategy Forum on Research Infrastructures) [2010]
http://ec.europa.eu/research/infrastructures/pdf/esfri/home/esfri_inspiring_excellence.pdf

ESFRI (European Strategy Forum on Research Infrastructures) European Roadmap for Research Infrastructures, Implementation Report 2009
http://www.europarl.europa.eu/meetdocs/2009_2014/documents/itre/dv/esfri_implementation_report_2009_/esfri_implementation_report_2009_en.pdf

ESFRI (European Strategy Forum on Research Infrastructures) European Roadmap for Research Infrastructures, Update 2008
http://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2008/esfri_roadmap_update_2008.pdf

ESFRI (European Strategy Forum on Research Infrastructures) European Roadmap for Research Infrastructures, Report 2006

## National Security

**Question 27: Are the identified emerging directions and research infrastructure capabilities for National Security right? Are there any missing or additional needed?**

Basic IT literacy will support national security and, as many HASS graduates will be in management positions in future, they should be trained in data management and curation issues in their undergraduate coursework, and then also in postgraduate work. This will reduce the risks to data security that occur when managers have no understanding of basic data structures.

**Question 28: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?**

[No answer provided]

**Question 29: Is there anything else that needs to be included or considered in the 2016 Roadmap for the National Security capability area?**

[No answer provided]

**Underpinning Research Infrastructure**

**Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?**

It must continually be emphasised that access to cultural collections is an indicator of the health of a society. The investment, for example, in commemorating WW1 recognises this and is to be welcomed, but there is much more to Australia's cultural history than WW1. As an example of how little information we have about Australian cultural heritage, consider the Indigenous languages and how little is recorded of each of them. Using AIATSIS figures (from Austlang) there are 474 Australian Indigenous languages with little or no resources available, 333 languages with some resources, 98 languages with a reasonable amount and just 45 languages with a good amount of information recorded. When it comes to languages of our region, for which Australia can act as a caretaker of language records, the situation is far worse. PARADISEC has been working to address this gap and to locate and digitise analog recordings created over the past 60 years.This is an example of underpinning current and future research into the diversity of languages and cultural expressions (music, song, oral traditions) in our region.

However, just digitising material is not enough. The digitisation has to be done to current standards, and there needs to be a repository that will ensure the longevity of the digitised objects, together with descriptive metadata.

**Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?**

PARADISEC is involved with the Open Language Archives Community (OLAC) and the Digital Endangered Languages and Musics Archives Network (DELAMAN), both international initiatives that link up similar international projects. OLAC in particular is a valuable service that harvests information and presents it in various formats about each of the world's languages.

**Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?**

PARADISEC should be listed as an 'Existing capability and infrastructure', as should AIATSIS for Australian languages. An emerging trend is the automated recognition of audio and video material which will result in transcription of events and speech similar to the way in which OCR can treat text. This will expand access to media recordings for cultural research and community access and is an enhancement of the digitisation capability.

**Data for Research and Discoverability**

**Question 33: Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?**

[No answer provided]

**Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?**

[No answer provided]

**Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?**

[No answer provided]

## Other comments

**If you believe that there are issues not addressed in this Issues Paper or the associated questions, please provide your comments under this heading noting the overall 20 page limit of submissions.**

[No answer provided]